G-04

多言語用例対訳作成のための手がかり用例提供システムの開発

Development of Starter Sentences Collection System to Create Multilingual Parallel-text

尾崎　俊 †　　福島　拓 †　　吉野　孝 ‡
Shun Ozaki　Taku Fukushima　Takashi Yoshino

1. はじめに

現在，在日外国人数の増加に伴い，多言語間コミュニケーションの機会が増加している．コミュニケーションを行う際，言語の違いは大きな障壁となる．一般に多言語の十分な習得は困難であり，言語の違いを克服するためには，機械翻訳のような技術支援が必要になる．しかし，医療分野のような生命に直接関係する業務では，十分な相互理解が得られなければ医療過誤に繋がる可能性がある．このため，医療現場ではコミュニケーションに極めて高い正確性が求められている [1]．多言語間での医療対話を支援するシステムでは，精度が保証された用例対訳を用いることが多い．

用例対訳の収集，提供を目的として，我々は医療分野を対象とした，多言語用例対訳共有システム TackPad の開発を行っている [2]．TackPad は人手での用例対訳の作成を行っている．人手による用例対訳の作成は，医療現場で求められている用例対訳を作成できる利点がある．しかし，以下のような問題を抱えている．

- 用例の作成場所は主に医療現場以外であるため，医療現場で必要な用例を想起しにくい．
- 用例対訳作成者の人数に限りがあるので，用例対訳の種類，数に限界がある．

人手による用例作成でなく，Web 上から多言語サイトを発見し，そのサイト内から用例対訳となる文を自動的に収集する研究が行われている [3]．本研究ではまず，単一言語の用例を人手によって作成するのを支援する．なお，本稿では日本語の用例作成を支援する．本研究の目的は，用例作成者の負担を減らすために，用例を作成する際に手がかりとなる文 (以下，手がかり用例とする) を Web 上から抽出することである．

本稿では，手がかり用例の検索手法および抽出手法を提案する．

2. 手がかり用例の抽出手法

Web 検索を用いた手がかり用例の抽出手法を，検索クエリの作成と抽出ルールに分けて説明する．

2.1 検索クエリの作成

本研究では，TackPad で用例が作成された時に用例作成者が付与できるタグと，“です” や “か” などの文の語尾を組み合わせたものを検索クエリとする．多くの用例対訳は口語であるため，語尾は “です” や “か” などの助動詞・助詞となっている．そのため，タグと語尾を組み合わせたものを検索クエリとした．

次に具体的な検索クエリの作成方法を示す．今回，タグを 5 つ，語尾を 3 つ用意し，それらを組み合わせた計 15 個を検索クエリとした．クエリの書式は「タグ+(空白)+語尾」である．具体的には「疲労　ます」のようなクエリを用いて検索を行う．

以下にタグと語尾の決定方法を示す．

(1) タグの選択

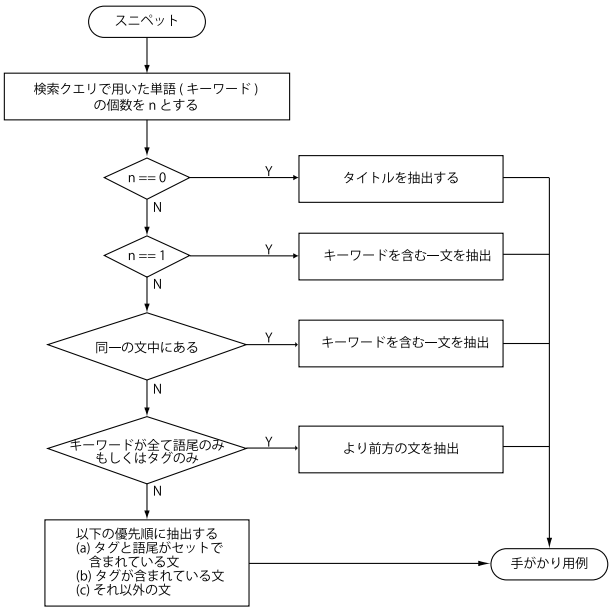

図 1: 抽出手法のフローチャート

TackPad で用例に付与されている全 287 個のタグから，症状に関係する 26 個のタグを抽出した．そのタグの中で使用頻度が低い 5 個のタグをランダムに抽出し，それらを検索で使用するタグとした．抽出したタグは「鬱」「震え」「疲労」「精神病」「ひび割れ」である．

(2) 語尾の選択

TackPad に登録されている全ての日本語用例を形態素解析によって品詞に分解し，その中から使用頻度が高い語尾上位 3 語を抽出した．形態素解析には Igo[*1] を使用した．抽出した語尾は「です」「ます」「ください」である．

2.2 手がかり用例の抽出

前節で述べた検索クエリを用いて Web 検索を行った．Web 検索には Google[*2] の検索エンジンを使用した．不適切な検索結果を減らすために，検索の設定をセーフサーチ強[*3] にして行った．検索結果 1 位から 10 位までのスニペット[*4] から次に示す抽出手法を適用して手がかり用例を抽出した．抽出手法のフローチャートを図 1 に示す．抽出手法の手順を以下に示す．

(1) 検索結果のスニペットから検索クエリで用いた単語 (以下キーワードとする) の数を数える．

(2) キーワードが一つもない場合は検索結果のページタイトルを抽出する．

† 和歌山大学大学院システム工学研究科，Graduate School of Systems Engineering, Wakayama University
‡ 和歌山大学システム工学部，Faculty of Systems Engineering, Wakayama University

[*1] http://igo.sourceforge.jp/
[*2] http://www.google.com/
[*3] Google 検索結果のページから，露骨な性描写を含むビデオや画像だけでなく，不適切なコンテンツにリンクされている可能性がある検索結果も除外される [4].
[*4] 検索された Web ページの要約文のこと．検索結果の一部として表示される．

表 1: 評価ごとの用例数

評価	1	2	3	4	合計
用例数	14	28	48	49	139
割合 (%)	10	20	35	35	100

評価 1：そのまま医療現場で使用できる
評価 2：一部を修正すると医療現場で使用できる
評価 3：大きな修正を行うと医療現場で使用できる
評価 4：全く医療現場で使用できない

(3) キーワードが一つの場合は，キーワードを含む一文を抽出する．

(4) 同じ文中にキーワードが複数ある場合はその文を抽出する．

(5) 異なる文中にキーワードがあり，キーワードの種類がタグのみ，もしくは語尾のみの場合，スニペット内のより前方にあるキーワードが含まれている文を抽出する．

(6) 異なる文中にキーワードがあり，キーワードの種類がタグと語尾の両方がスニペット中に含まれている場合，以下の条件順に当てはまる文を抽出する．

  (a) 検索クエリで使用したタグと語尾がセットで含まれている文

  (b) 検索クエリで使用したタグが含まれている文

  (c) それ以外の文

3. 実験

抽出した 150 文 (「症状タグ 5 種類」×「語尾 3 種類」×「検索結果上位 10 件」) から重複した 11 文を除く，139 文の評価を大学生 3 名に依頼した．評価基準は以下の 4 段階とした．

評価 1 そのまま医療現場で使用できる

評価 2 一部を修正すると医療現場で使用できる

評価 3 大きな修正を行うと医療現場で使用できる

評価 4 全く医療現場で使用できない

各被験者はランダムに並び変えられた 139 文を上記の評価基準に従って評価した．

4. 実験結果と考察

本章では，実験結果と考察を述べる．なお，本稿では各用例の評価の中央値を，その用例の評価とする．表 1 に評価ごとの用例数を示す．以降の考察では，評価 1(そのまま医療現場で使用できる) と，評価 4(全く医療現場で使用できない) の特徴を，検索クエリで使用した「タグ」「語尾」に分けて考察を行う

4.1 タグとの関係

検索クエリとして使用したタグと，用例の評価の関係を表 2 に示す.「ひび割れ」「震え」を検索クエリとして抽出した用例は，評価 4(全く医療現場で使用できない) に多く分類された．これらのタグは「人以外を主語にとることができる」という特徴がある．実験で抽出された例としては「天井から窓枠まで、クロスの継ぎ目を一直線にひび割れ」が当てはまる．

このため，人のみを主語にとることができる単語を検索クエリとして使用することで，より医療現場で使用される用例を検索できる可能性がある．

4.2 語尾との関係

検索クエリとして使用した語尾と，用例の評価の関係を表 3 に示す.「ください」と「ます」の語尾の数に特徴があった．以降,「ください」を検索クエリとして使用した時の評価別の割合を用いて考察を行う．評価 1(そのまま医療現場で使用できる) に分割された割合は 14%であった．また，評価 4(全く医療現場で使用できない) に分割された割合は 43%と，評価 1 よりも多くみられた．

表 2: タグと用例の評価との関係

		疲労	鬱	精神病	震え	ひび割れ	合計
評価 1	用例数	5	4	3	1	1	14
	割合 (%)	36	29	21	7	7	100
評価 4	用例数	7	14	4	11	13	49
	割合 (%)	14	29	8	22	27	100

評価 1：そのまま医療現場で使用できる
評価 4：全く医療現場で使用できない

表 3: 語尾と用例の評価との関係

		ください	ます	です	合計
評価 1	用例数	2	9	3	14
	割合 (%)	14	64	21	100
評価 4	用例数	21	15	13	49
	割合 (%)	43	31	27	100

評価 1：そのまま医療現場で使用できる
評価 4：全く医療現場で使用できない

この原因として,「ください」には，語尾に付けることで相手に何かを要望・懇願する意を表すものと,「くれ」の尊敬語の二通りの意味がある．今回の実験で評価 4 の用例の中には「くれ」を主の意味としたものがあった．実験で抽出された例としては「死ぬほど鬱になる画像ください」「外壁のひび割れについてアドバイスください」が当てはまる．

これらのことから，検索クエリとして使用した語尾の品詞が，抽出した用例内で異なる品詞で使用されている用例は評価が低いことが分かった．そのためそれらの用例は抽出から除外することで，さらに抽出手法の精度が上がる可能性がある．

5. おわりに

本稿では，Web 検索を用いて用例対訳作成のための手がかりとなる用例を抽出する手法を提案し，抽出した用例が医療現場で使用できるかどうかの評価を行った．

本研究の貢献は次の 2 つである．

(1) 本手法を用いることで，約 10%の割合で「そのまま医療現場で使用できる用例」を Web 上から抽出できることを示した．また，約 65%の割合で「医療現場で使用する用例を作成する手がかりとなる用例」を Web 上から抽出できることを示した．

(2) 検索クエリとして使用した語尾の品詞が，抽出した用例内で異なる品詞で使用されている用例は評価が低い．

今後，本手法をシステムに適用し，さらに実験を重ねて改良していく．また，今回は日本語の用例の抽出を行ったが，今後多言語への適用を目指す．

謝辞

本研究の一部は，日本学術振興会科学研究費 基盤研究 (B) (22300044) の補助を受けた．

参考文献

[1] 田村太郎，多民族共生社会ニッポンとボランティア活動，明石書店，2000.

[2] 福島　拓，宮部真衣，吉野　孝，重野亜久里：医療分野を対象とした多言語用例対訳収集 Web システム TackPad の開発，マルチメディア，分散，協調とモバイル（DICOMO2008）シンポジウム，pp.1030–1036 (2008).

[3] Jisong Chen, Rowena Chau, Chung-Hsing Yeh :Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32 (2004).

[4] セーフサーチとは-ウェブマスターツールヘルプ：http://www.google.com/support/webmasters/bin/answer.py?hl=ja&answer=510